Measuring the Biases and Effectiveness of Content-Style Disentanglement

Xiao Liu [*,1]
Xiao.Liu@ed.ac.uk

Spyridon Thermos [*,1]
SThermos@ed.ac.uk

Gabriele Valvano [*,1,2]
gabriele.valvano@imtlucca.it

Agisilaos Chartsias[1]
Agis.Chartsias@ed.ac.uk

Alison O'Neil [1,3]
Alison.ONeil@mre.medical.canon

Sotirios A. Tsaftaris [1,4]
S.Tsaftaris@ed.ac.uk

[1] School of Engineering
University of Edinburgh
Edinburgh, UK

[2] IMT School for Advanced Studies Lucca
Lucca, Italy

[3] Canon Medical Research Europe Ltd.
Edinburgh, UK

[4] The Alan Turing Institute
London, UK

[*] Equal contribution

Abstract

A recent spate of state-of-the-art semi- and un-supervised solutions disentangle and encode image "content" into a spatial tensor and image appearance or "style" into a vector, to achieve good performance in spatially equivariant tasks (*e.g.* image-to-image translation). To achieve this, they employ different model design, learning objective, and data biases. While considerable effort has been made to measure disentanglement in vector representations, and assess its impact on task performance, such analysis for (spatial) content - style disentanglement is lacking. In this paper, we conduct an empirical study to investigate the role of different biases in content-style disentanglement settings and unveil the relationship between the degree of disentanglement and task performance. In particular, we consider the setting where we: (i) identify key design choices and learning constraints for three popular content-style disentanglement models; (ii) relax or remove such constraints in an ablation fashion; and (iii) use two metrics to measure the degree of disentanglement and assess its effect on each task performance. Our experiments reveal that there is a "sweet spot" between disentanglement, task performance and - surprisingly – content interpretability, suggesting that blindly forcing for higher disentanglement can hurt model performance and content factors semanticness. Our findings, as well as the used task-independent metrics, can be used to guide the design and selection of new models for tasks where content-style representations are useful. Code is available at `https://github.com/vios-s/CSDisentanglement_Metrics_Library`.

1 Introduction

Recent work in representation learning argues that to achieve explainable and compact representations, one should separate out, or *disentangle*, the underlying explanatory factors into different dimensions of the considered latent space [2, 24]. In other words, it is beneficial

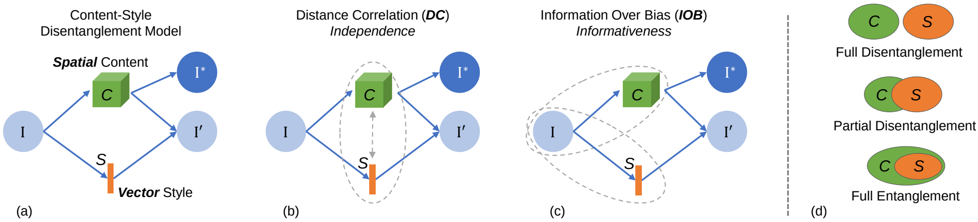

Figure 1: **(a)** A schematic representation of disentanglement between spatial content C and vector style S in the context of a primary and a secondary spatially equivariant task (I′, I*). Measuring the degree of C-S disentanglement using distance correlation **(b)** and information encoded over the input bias **(c)**. **(d)** A visual description of degrees of C-S (dis)entanglement.

to obtain representations that can separate latent variables that capture sensitive and useful information for the task at hand from the less informative ones [1]. Disentanglement has recently shown to improve task performance, model generalization, and representation interpretability [11, 15, 19, 30, 41, 43, 48, 57]. Unfortunately, disentangling without supervision is an ill-posed and impossible task [36, 37, 52] and, to obtain it, we must introduce restrictions and inductive priors [36, 37]. These priors are different forms of "bias" imposed by model design (design bias), learning objectives (learning bias), and data (data bias).

In this work, we set out to reveal such choices of bias in state-of-the-art (SoTA) disentanglement methods. Our particular focus is on "content-style" disentanglement, which decomposes input images into spatial "content" and vector "style" representations. In principle, content variables (C) should contain the semantic information required for spatially equivariant tasks (*e.g.* segmentation and pose estimation), whereas style variables (S) contain information on image appearance (*e.g.* colour intensity and texture). However, contrary to extensive research on quantifying the degree of disentanglement between vectors [9, 17, 18, 28, 31, 44, 55], usually there is no analysis of C-S disentanglement. In fact, to the best of our knowledge, there is no study identifying the training biases enforced in C-S disentanglement settings or exposing the true relationship between the degree of disentanglement and model performance. Herein, we attempt to bridge these gaps with our **contributions**:

- We identify and analyse the key biases in SoTA models that employ C-S disentanglement. We show how the biases affect disentanglement and task performance (utility) in three popular vision tasks: image translation, segmentation, and pose estimation.
- To make a quantitative analysis possible, we propose two complementary metrics building on existing work, to evaluate C-S disentanglement (Fig. 1) in terms of amount of information encoded in each latent variable (informativeness) and (un)correlation between the encoded *spatial tensor* content and *vector* style (proxy for independence).
- We find that: a) lower C-S disentanglement benefits task performance if a specific style-related prior is not violated; and b) performance is highly correlated with latent variable informativeness. We also assess content semanticness (interpretability).

2 Related Work

Content-Style Disentanglement. Image-to-Image translation has extensively explored the decoupling of image style and content [27, 32, 33, 34]. Content-style disentanglement was also used in other applications, such as semantic segmentation [8] and pose estimation [7], where the content serves as a robust representation for downstream tasks. In general, most methods derive latent spaces capturing C or S information using auto-encoder variants.

These models achieve C-S disentanglement through different biases, such as architectural choices (*e.g.* AdaIN [26], content binarization [8]), learning objectives (*e.g.* Kullback-Leibler divergence, latent regression loss, de-correlation losses in vector representations [6, 49]), or supervisory signals (*e.g.* using content for segmentation [8]). However, the precise effect of each bias on disentanglement and model performance is not thoroughly explored.

Evaluating Disentanglement. Recently, several methods have been proposed for assessing the degree of disentanglement in a vector latent variable. A classical approach is *latent traversals*: a visualization showing how traversing single latent dimensions generates variations in the image reconstruction. Latent traversals do not need ground truth information on the factors, and can be used in mixed tensor spaces [8, 38] to offer qualitative evaluations. Alternatively, latent traversals can be combined with pre-trained networks to measure the perceptual distance between the produced embeddings [28].

There exist several ways in quantitatively evaluating representations learned by VAEs and GANs. Unfortunately, these methods rely only on vector representations, and some also peruse ground truth knowledge about the latent factors. In particular, some methods try to associate known factors of variations (*e.g.* rotation) with specific latent dimensions [23, 29] or manifold topology [59]. Others measure the ability to isolate one factor in a single vector latent variable [31], measuring compactness or modularity [9, 18, 55], linear separability [28], consistency and restrictiveness [47], and explicitness of the representation [44]. Lastly, there is work on measuring the factor informativeness in a vector latent variable *w.r.t.* the input, independence among factors, as well as interpretability [17, 18].

The aforementioned metrics cannot be directly employed to C-S disentanglement settings, where the latent factors have different dimensionality (*i.e.* the style is a vector and the content a spatial multi-channel tensor). However, in this paper we attempt to transfer these concepts to the C-S disentanglement domain, incorporating both spatial (tensor) and vector representations[1] to expand our understanding of the relation between C-S disentanglement and: a) biases adopted by each model; b) task performance; c) representation interpretability.

3 Measuring Properties of Disentangled Content and Style

Given N image samples $\{I_i\}_{i=1}^N$, we assume two representations of content and style: $\{C_i\}_{i=1}^N$ and $\{\underline{s}_i\}_{i=1}^N$, respectively. Building on existing work in vector-based disentanglement [17, 18], we present two complementary metrics to evaluate two properties in the context of C-S disentanglement: *(un)correlation*, and *informativeness*. We provide evidence that the metrics offer complementary information in supplement Sec. 8. Then, we discuss two properties of the disentangled representations, namely their *utility* and *interpretability*.

Distance Correlation (*DC*). Disentangled representations separate content and style into independent latent spaces [24], satisfying $p(C,\underline{s}) = p(C)p(\underline{s})$. However directly measuring independence between spatial C and vector S with existing metrics is not feasible. Since independent representations must be uncorrelated [9], we use the *empirical Distance Correlation* (*DC*) [51] to measure the correlation between tensors of arbitrary dimensionality. Note that *DC* is bounded in the $[0,1]$ range, while differently from other correlation-independence

[1] Note that the metrics used for our analysis are generic and can be readily applied to vector-based C-S disentanglement methods, such as [20].

metrics, such as the kernel target alignment [13] and the Hilbert-Schmidt independence criterion [21], it has the advantage of not requiring any pre-defined kernels.

For N samples, consider two N-row matrices T_1 and T_2. In general, T_1 and T_2 row dimension varies as they are formed by concatenating images I_i, content features C_i or style features $\underline{s}_i$. For I_i and C_i we first concatenate the channels and then row-scan to form a vector; $\underline{s}_i$ is already a vector. DC is then defined as:

$$DC(T_1,T_2) = \frac{dCov(T_1,T_2)}{\sqrt{dCov(T_1,T_1)dCov(T_2,T_2)}}, \text{ with } dCov(X,Y) = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{A_{i,j}B_{i,j}}{N^2}}. \quad (1)$$

Here, $dCov$ is the distance covariance between any two N-row matrices X and Y, while A and B are their respective distance matrices. In particular, each matrix element $a_{i,j}$ of A is the Euclidean distance between two samples $||X^i - X^j||$, after subtracting the mean of row i and column j, as well as the matrix mean. B is similarly calculated for Y. We estimate disentanglement between C and S using distance correlation, $DC(C,\underline{s})$, with values closer to 0 indicating higher disentanglement. C and S can be uncorrelated, *e.g.* $DC(C,\underline{s}) = 0$, either when they encode unrelated information or when one encodes *all* information and the other encodes *noise*. The latter indicates posterior collapse, thus full entanglement. To tackle this, $DC(C,\underline{s})$ needs a complementary metric to measure the representations' informativeness.

Information Over Bias (*IOB*). To explicitly measure the amount of information encoded in C and S, we introduce the *Information Over Bias* (*IOB*) metric, aiming to detect posterior collapse when C and S are disentangled, but one (C or S) is not informative about the input. Given latent variables $z \in \{C,\underline{s}\}$ produced from N images at inference, we measure the amount of information encoded in each representation. To do so, we train a decoder G_{θ_l}, a neural network with parameters θ_l, to reconstruct images I, given the features z.

Thus, we define IOB as the expectation over the test images of the ratio:

$$IOB(I,z) = \mathop{\mathbb{E}}_{i}\left[\frac{\text{MSE}(I_i, G_{\theta_1}(\mathbb{1}))}{\text{MSE}(I_i, G_{\theta_2}(z_i))}\right] = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\frac{1}{K}\sum_{k=1}^{K}||I_i^k - G_{\theta_1}(\mathbb{1})||^2}{\frac{1}{K}\sum_{k=1}^{K}||I_i^k - \tilde{I}_i^k||^2 + \varepsilon}\right), \quad (2)$$

where I and $\tilde{I}$ are an image and its reconstruction obtained through G_{θ_l}; $i = 1\ldots N$, $k = 1\ldots K$, $l = 1\cdots+\infty$ are indices iterating on the test images, the image pixels, and the generator model index (different for each run), respectively; ε is a small value that prevents division by zero. We justify the above definition of IOB by observing that a post-hoc minimization of the MSE between $\tilde{I}$ and I is equivalent to maximizing the log likelihood (see our analysis in supplement Sec. 1). Note that the ratio aims at ruling out from IOB both data correlations (common structure, colours, pose, etc., across the images of the dataset) and architectural biases that one could introduce in the design of G_{θ_l}. In particular, this is done by computing the ratio between the MSE obtained after training G_{θ_l} to reconstruct the images from their *informative* representation z (*i.e.* $\text{MSE}(I_i, G_{\theta_2}(z_i))$), and after training G_{θ_l} from an *uninformative* constant tensor $\mathbb{1}$ (*i.e.* $\text{MSE}(I_i, G_{\theta_1}(\mathbb{1}))$). In the latter case, G_{θ_l} will only learn the dataset bias it can model, given θ_l. Hence, high values of IOB can be associated with higher information inside the representation z, while the lower bound $IOB = 1$ means that no information of the images I is encoded in z.[2]

[2] Optimising G_θ with stochastic gradient descent can introduce noise and slightly alter the measure. For example, IOB may, in practice, even be slightly smaller than 1. Thus, we average results across multiple runs and initializations of G_θ, which contributes to the computational load of estimating IOB.

Utility and Interpretability. As discussed, we can use *DC* and *IOB* to measure the degree of disentanglement between latent representations. However, one of the primary goals of disentanglement is to improve task performance (utility) and representation interpretability, hence we also investigate the relationship between C-S disentanglement and these two notions. In particular, we measure utility by quantifying performance on a downstream task, which for disentangled representations is typically image translation [27, 33] to translate image content from one domain to another. We also consider tasks using content *e.g.* to extract segmentations [8] or landmarks [38], and therefore assess how effectively it can be used in downstream tasks. We detail performance metrics for each application in Sec. 5.

Assessing interpretability is not trivial. Here, we assume that interpretability implies semantic representations. Previously, vector representations were considered semantic if a portion of the latent space corresponded to specific data variations [10, 60]. Style semantics were qualitatively evaluated with latent traversals of individual dimensions [8]. Thus, we consider a style interpretable if images produced by linear traversals in the style latent space are realistic and smoothly change intensity. In spatial representations, such data variation should be confined to individual objects: thus, semantic content should split distinct objects into separate channels of C. Wherever possible, we evaluate this with qualitative visuals.

4 Validating the Effectiveness of *DC* and *IOB*

To verify the effectiveness of *DC* and *IOB*, we design an experiment using the synthetic teapot dataset [18], which consists of 200k of 64×64 pixel resolution images of a teapot with varying pose and colour. Each image of this dataset is generated using 5 ground truth (GT) generating factors, *i.e.* azimuth, elevation, red, green, and blue colour, independently sampled from 5 uniform distributions. We consider the 3 colour factors as the GT style (GT S) representation, while as GT spatial content (GT C) we leverage the object's segmentation mask, as it correlates with the azimuth and elevation factors (see Sec. 2 of the supplement).

We first evaluate *DC* and *IOB* using the GT C and S representations, and the input images. Then, we sample from a uniform distribution $U[0,1]$ to generate a random style and content representations for each image, and evaluate the metrics using the following scenarios: a) random C, GT S and images; b) GT C, random S and images; c) random C, random S and images. Finally, to approximate the highly entangled C and S scenario, we construct the content-correlated style representations (correlated S) as the azimuth, elevation and red colour factors. For each experiment, we randomly sample 5k images and the GT representations, while all results are the average of 3 different runs.

Results. From Table 1, we observe that for any combination of C and S (except for the correlated S one), the $DC(C,\underline{s})$ is low, which indicates that the representations are highly

Table 1: Empirical study results for the *DC* and *IOB* metrics evaluation using the teapot dataset [18]. Results are in “mean ±std" format.

Metric	GT C	Random C	GT C	Random C	GT C
	GT $\underline{s}$	GT $\underline{s}$	Random $\underline{s}$	Random $\underline{s}$	Correlated $\underline{s}$
$DC(C,\underline{s})$ (↓)	0.17 ±0.00	0.13 ±0.04	0.05 ±0.00	0.13 ±0.04	0.53±0.02
$DC(I,C)$ (↑)	0.64 ±0.03	0.16 ±0.05	0.64 ±0.03	0.16 ±0.05	0.64±0.03
$DC(I,\underline{s})$ (↑)	0.87 ±0.00	0.87 ±0.00	0.04 ±0.00	0.04 ±0.00	0.33±0.00
$IOB(I,C)$ (↑)	1.73 ±0.10	1.41 ±0.20	1.73 ±0.10	1.41 ±0.20	1.73±0.10
$IOB(I,\underline{s})$ (↑)	2.47 ±0.78	2.47 ±0.78	0.76 ±0.15	0.76 ±0.15	2.70±0.26

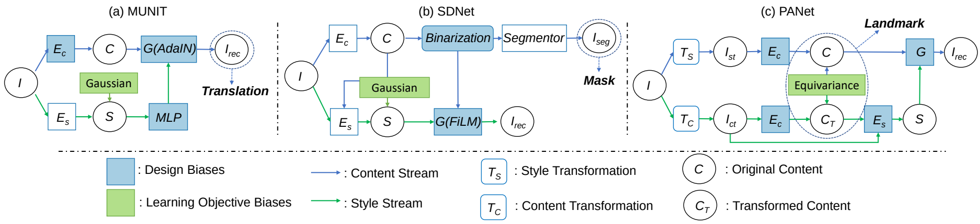

Figure 2: Model schematics. a) MUNIT: Instance normalization is used to remove style from content; E_s uses global pooling. b) SDNet: the content is represented with binary features; style is forced to approximate a normal prior. c) PANet: content and style are encouraged to be equivariant to intensity and spatial transformations.

uncorrelated. This result meets our expectation as the colour (S) and the azimuth or the elevation factors (C) are independent in the teapot dataset. However, we also observe a high $DC(C,\underline{s})$ value, *i.e.* 0.53, between GT C and correlated S, which verifies that DC can indeed detect the entangled representations case. Additionally, the effectiveness of the DC metric is validated by the high $DC(I,C)$ values when using GT C representations, versus the low values when using random C ones. Note that the DC between the GT S and image is higher than the one between GT C and image, which is reasonable as S and image have nearly one-to-one mapping relationship, while the segmentation masks for different images can be similar. The IOB results, reported in Table 1, also reflect that the segmentation mask is less informative ($IOB(I,C) = 1.73$) about the input image compared to S ($IOB(I,\underline{s}) = 2.47$) for the GT C and GT S case. This is a result of the strong dataset bias, where given that the object is always a teapot, it is the colour of the reconstructed image that makes it more similar to the input one in terms of MSE.

5 Experimenting on Vision and Medical Applications

Many applications disentangle C from S [4, 20, 40, 46] or other attributes, such as pose, geometry, and motion [14, 25, 54, 56], to improve performance in vision tasks. For our analysis, we select and discuss three SoTA approaches (see Fig. 2) from diverse applications, namely image translation (MUNIT [27]), semantic segmentation (SDNet [8]), and pose estimation (PANet [38]). All resemble auto-encoders, mapping input images to disentangled features but use several biases, which are detailed below. Our scope is to elucidate how each bias affects disentanglement using these models and their chosen biases as exemplars.

Here we describe how each bias is enforced, whilst the detailed model descriptions and a summary of their *design* and *learning* biases can be found in Sec. 4 of the supplement. In particular, for: **a) MUNIT** we consider ablations removing Instance Normalization (IN) [53], AdaIN layers, or style Latent Regression (LR) loss (for fairness, we do not remove LR of the content as it is fundamental for the functioning of the model); **b) SDNet** we identify content binarization, Gaussian approximation, LR and the FiLM-based [42] decoder as the main biases that affect C-S disentanglement. We investigate their impact on the representations and their effect on semantic segmentation; **c) PANet** we remove the Gaussian prior and replace its specific C-S conditioning with AdaIN. We analyse PANet performance in pose estimation. These models help us cover the following diverse cases: **i)** no supervision and weak C constraints (MUNIT), **ii)** no supervision with strong C constraints (PANet), and **iii)** supervision with strong C constraints (SDNet).

Table 2: Comparative evaluation of MUNIT variants using the proposed metrics. We use *FID* and *LPIPS* to measure translation quality and diversity between SYNTHIA [45] and Cityscapes [12] samples. Results are in “mean ±std" format.

		Learning Bias	**Design Bias**	
Metric	Original Model	w/o Latent Regression (LR)	w/o AdaIN	w/o Instance Normalization (IN)
$DC(C,\underline{s})$ (↓)	0.44 ±0.06	**0.40** ±0.08	0.43 ±0.01	0.66 ±0.03
$DC(I,C)$ (↑)	0.57 ±0.07	0.57 ±0.08	0.58 ±0.08	**0.73** ±0.03
$DC(I,\underline{s})$ (↑)	0.70 ±0.02	**0.73** ±0.03	0.56 ±0.03	0.63 ±0.05
$IOB(I,C)$ (↑)	4.36 ±0.38	4.34 ±0.58	4.85 ±0.10	**5.01** ±0.12
$IOB(I,\underline{s})$ (↑)	1.31 ±0.04	**1.46** ±0.05	1.17 ±0.04	1.28 ±0.06
FID (↓)	73.48 ±8.35	104.51 ±4.21	**52.48** ±5.03	71.4 ±4.86
LPIPS (↑)	0.08 ±0.01	0.09 ±0.01	0.06 ±0.01	**0.10** ±0.01

Setup. For each model, we analyze the effect that design choices and learning objectives have on disentanglement and task performance, and we evaluate utility and interpretability of the learned representations. We use the implementations provided by the authors, ablating only the components needed for our analysis. In all tables, arrows (↑,↓) indicate direction of metric improvement; best results are in bold. Numbers are the average of 5 different runs. Data description and learning settings can be found in supplement Sec. 4 (see D.1-D.4).

5.1 Image-to-Image Translation

We consider the original MUNIT and three variants: **i)** we replace the AdaIN modules of the decoder with simple style concatenations, reducing the restrictions on the re-combination of C and S. **ii)** We remove the LR loss, responsible for the style following a Gaussian. **iii)** We remove IN from the content encoder, to confirm that it helps to cancel out original style and retain the content only [26]. As [27] we evaluate quality and diversity of the translated images using the Fréchet Inception Distance (*FID*) [22] and LPIPS [58].

Results. Table 2 reports the results of the ablations on the SYNTHIA [45] and Cityscapes [12] datasets. Replacing AdaIN (**w/o AdaIN**) with simple concatenation does not affect the level of C-S disentanglement, but it leads to a 0.14 absolute decrease in $IOB(I,\underline{s})$ and $DC(I,\underline{s})$, indicating that the style becomes less informative and less correlated with the input. Here, we observe an information shift to the content (lower $IOB(I,\underline{s})$, higher $IOB(I,C)$) leading to better translation quality but worse diversity ($LPIPS = 0.06$). We infer that this variant is worse than the original model, which had more balanced quality/diversity scores. By removing the LR learning bias (**w/o LR**), the style becomes more correlated to the input image. If the style distribution is no longer Gaussian, the style has more degrees of freedom to encode non-relevant information, which contributes to higher $IOB(I,\underline{s})$ and higher C-S disentanglement. This ablation leads to a significant translation quality decrease, while contrary to the analysis in [27], the diversity is not negatively affected. Finally, by removing IN (**w/o IN**) we expect a more entangled content that is encoding also some style information. Our expectations are confirmed by the decrease in C-S disentanglement ($DC(C,S) = 0.66$), and a more informative content (which is also more correlated to the input image). Interestingly, relaxing the content constraints for a task that does not require a strictly semantic content (such as image segmentation), leads to the best quality/diversity balance. Note that we define the best balance as achieving the highest average ranking in *FID* and *LPIPS* (*e.g.* the “w/o IN" model variant is the 1st in *LPIPS* and 2nd in *FID*).

Summary. Our experiments reveal a trade-off between the translation quality/diversity

Table 3: Comparative evaluation of SDNet variants using the proposed metrics. We use the *Dice* score to measure semantic segmentation performance on the ACDC [3] dataset with 1.5% annotation masks. Results are in "mean ±std" format.

		Learning Bias	Design Bias	
Metric	Original Model	w/o KLD and Latent Reg. (LR)	w/o Binarization	SPADE
$DC(C,\underline{s})$ (↓)	0.49 ±0.02	0.64 ±0.03	**0.44** ±0.00	0.52 ±0.01
$DC(I,C)$ (↑)	0.94 ±0.01	0.94 ±0.01	**0.98** ±0.02	0.93 ±0.01
$DC(I,\underline{s})$ (↑)	0.43 ±0.02	**0.66** ±0.00	0.44 ±0.01	0.45 ±0.01
$IOB(I,C)$ (↑)	4.71 ±0.26	4.84 ±0.23	**5.89** ±0.22	5.09 ±0.00
$IOB(I,\underline{s})$ (↑)	1.00 ±0.01	1.00 ±0.04	0.98 ±0.04	1.00 ±0.04
Dice (↑)	0.62 ±0.02	0.61 ±0.04	0.63 ±0.04	**0.75** ±0.02

and disentanglement in a translation task.[3] Our metrics indicate that a partially disentangled C-S space –with a near-Gaussian style latent space– leads to the best quality/diversity performance. For MUNIT this is achieved by removing the IN design bias.

5.2 Medical Segmentation

In SDNet, content binarization and style Gaussianity are the key representation constraints. We evaluate their effect and those of decoder design on segmentation performance measuring the Dice Score [16, 50] after: **i)** removing content thresholding (w/o Binarization), **ii)** removing style Gaussianity (w/o Kullback-Liebler Divergence (KLD) and LR), and **iii)** considering a new decoder, obtained replacing the FiLM style conditioning with *SPADE* [39]. SPADE is less restrictive, allowing the style to encode more image-related information, such as textures, rather than just intensity (see supplement Sec. 6.1).

Results. Table 3 reports our findings on the ACDC [3] dataset. We highlight that when using all the available annotations (fully supervised learning), all SDNet variants achieve a similar accuracy (see supplement Sec. 6.2 for more details), suggesting that strong learning biases, such as supervised segmentation costs, make disentanglement less important. Thus, we consider the semi-supervised training case with minimal supervision, using only the 1.5% of available labelled data. Overall, the style encodes little information in all SDNet variants, probably because all medical images in ACDC have similar styles (data bias), and reconstructing using an average style is enough to have low $IOB(I,S)$. However, C-S disentanglement is still important to obtain a good content representation. For example, intermediate levels of disentanglement (**SPADE**) lead to the best segmentation performance. In this variant, disentanglement decreases compared to the original model, as some style information is probably leaked to the content (higher $DC(C,\underline{s})$ and $IOB(I,C)$). On the other hand, also removing C binarization (**w/o Binarization**) makes content more informative; since the correlation between C and S decreases, we assume that the extra information encoded in C is not part of the style. Lastly, removing the Gaussian prior constraints from the style (**w/o KLD and LR**) leads to the lowest degree of disentanglement as there is no information bottleneck on S, and a slight decrease of the Dice score.

Summary. We find disentanglement to have minimal effect on task performance when training with strong learning signals (*i.e.* supervised costs). In the semi-supervised setting, a higher (but not full) degree of disentanglement leads to better performance, while the amount of information in C alone is not enough to achieve adequate segmentation performance.

[3] Note that the effect of C-S disentanglement on task performance also depends on the data bias.

Table 4: Comparative evaluation of PANet variants using the proposed metrics. We use *SIM* to measure the performance in terms of pose estimation from landmarks on the DeepFashion [35] dataset. Results are in “mean ±std" format.

		Learning Bias	Design Bias		
Metric	Original Model	w/o Equivar.	AdaIN w/o Gaussian	AdaIN	MLP
$DC(C,\underline{s})$ (↓)	0.65 ±0.01	0.76 ±0.08	**0.25** ±0.01	0.36 ±0.02	0.69 ±0.03
$DC(I,C)$ (↑)	0.59 ±0.01	**0.60** ±0.02	0.53 ±0.01	0.56 ±0.01	0.58 ±0.02
$DC(I,\underline{s})$ (↑)	**0.83** ±0.01	0.82 ±0.01	0.38 ±0.06	0.81 ±0.01	0.82 ±0.03
$IOB(I,C)$ (↑)	1.50 ±0.08	1.50 ±0.08	**1.53** ±0.06	1.52 ±0.08	1.49 ±0.06
$IOB(I,\underline{s})$ (↑)	1.09 ±0.04	1.13 ±0.06	1.12 ±0.09	1.10 ±0.15	**1.21** ±0.09
SIM (↑)	**0.71** ±0.02	0.47 ±0.04	0.58 ±0.00	0.64 ±0.01	0.68 ±0.01

5.3 Pose Estimation

We consider the original PANet model and four possible variants, relaxing design biases on both C and style, and learning biases. In detail: **i)** we experiment with a different conditioning mechanism to re-entangle S and C, that consists of the use of AdaIN, rather than just multiplying each S vector with a separate C channel (introducing a bias on S, similar to MUNIT). **ii)** We consider the case where, instead of learning a different S for each channel of C, we extract a global S vector, predicted by an MLP (relaxing the tight 1:1 correspondence between C and S channels). **iii)** We also consider the case where each C part is not approximated by a Gaussian prior. Since we cannot use the original decoder to combine C and S, we reintroduce S using AdaIN. **iv)** Finally, we evaluated the effect of the equivariance constraint, by removing it from the cost function.

Results. Table 4 reports results of the ablations on the DeepFashion [35] dataset. We assess model performance using *SIM* [5] to measure the similarity between the predicted and ground truth landmarks visualized as heatmaps. Whilst the original model is the best to predict landmarks, it only achieves average disentanglement (see $DC(C,\underline{s})$). Using an **AdaIN**-based decoder consistently improves disentanglement as it has a strong inductive bias on the re-entangled representation (see $DC(C,\underline{s})$ for **AdaIN**, and **AdaIN w/o Gaussian**), but it leads to worse landmark detection – the representation adapts tightly to the strongly-biased decoder, and the content loses transferability to other tasks, and interpretability (see Figs. 3 and Fig. 7 of supplement). Using an **MLP** to encode S relaxes the specific conditioning between C and S (design bias) and reduces disentanglement. There is an information shift from C to S, as indicated by the higher $IOB(I,\underline{s})$ and the high $DC(C,\underline{s})$. Here, a moderate decrease of disentanglement shows slightly lower task performance. Finally, the equivariance cost is the most important factor for disentanglement; removing it (**w/o Equivariance**) leads to the most entangled representation (high $DC(C,\underline{s})$), and accuracy decrease in landmark detection.

Summary. Overall, lowering disentanglement leads to better landmark detection. Again, balance is the key to improve the auxiliary tasks. Here, partial disentanglement is achieved by carefully balancing the design biases used to extract style and to reintroduce it to content while decoding. Relaxing such biases with AdaIN or MLP makes landmark detection worse.

5.4 Discussion

We now discuss the relationship between C-S disentanglement and inductive biases, task performance, interpretability of the latent representations.

Do biases affect C-S disentanglement? Results in Sec. 5 illustrate that learning and design biases critically affect disentanglement. However, no evaluation can specifically char-

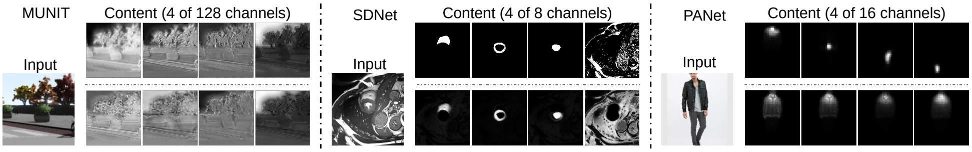
MUNIT
Content (4 of 128 channels)
Input
SDNet
Content (4 of 8 channels)
Input
PANet
Content (4 of 16 channels)
Input

Figure 3: Content *interpretability* of each original model (top row) and a variant with the most correlated C and S (bottom row). Removing content-related design biases from SDNet and PANet leads to less interpretable representations (same objects/joints appear in different channels), such as the unconstrained content representations of MUNIT.

acterize the relative importance of each one, since this depends on the task at hand, as well as the utilized data. In MUNIT, disentanglement is mainly encouraged by the content-related design and learning biases. In fact, IN is key to removing style information from the content, and the model cannot be successfully trained without LR of the content. Disentanglement in SDNet is susceptible to the biases that affect both latent variables. Using a SPADE decoder or removing content thresholding leads to more entanglement, while making the style Gaussian through learning constraints restricts its informativeness and encourages disentanglement. Similarly, PANet disentanglement is affected both by designing the content as Gaussian, and by the equivariance of C and S *w.r.t.* spatial or intensity transformations, respectively.

What is the relationship between C-S disentanglement and task performance? Our results showcase a clear sweet spot between C-S disentanglement and downstream task performance. In particular, we observe that lowering disentanglement by relaxing constraints on the content (*e.g.* removing IN), but preserving the biases that enforce style priors, such as C-S equivariance, leads to better performance.

Does disentanglement affect content interpretability? Interpretability is hard to quantify without metrics. Here, we consider the C interpretable if distinct objects appear in different channels. We qualitatively analyze C interpretability in Fig. 3 (see also supplement Sec. 9). Interpretability varies a lot with different *design biases* of the model, while learning biases do not seem to affect it. Without restrictive design bottlenecks on C, MUNIT spreads the content across channels. Instead, SDNet and PANet original models encourage C to encode different objects, or parts, into different channels. In SDNet, a semantic content is encouraged by applying Softmax across channels and then binarize the output features, while PANet approximates body parts as 2D Gaussians enforcing an information bottleneck on each channel of C. Removing the C constraints from SDNet and PANet spreads the spatial information across all channels, decreasing interpretability.

6 Conclusion

In this paper we evaluated the disentanglement between image C and S through experimenting on 3 SoTA models, and showcased how design and learning biases affect disentanglement and by extension task performance. Our findings suggest that whilst content-style disentanglement enables the implementation of certain equivariant tasks, partially (dis)entangled can lead to better performance than fully disentangled ones. Additionally, our analysis suggests that strict design constraints on the content space lead to increased interpretability, which could be exploited in post-hoc tasks. Using our findings and the presented metrics will enable the design of better models that achieve the degree of disentanglement that maximizes performance, rather than blindly pursuing very high (or low) disentanglement.

7 Acknowledgement

This work was supported by the University of Edinburgh, the Royal Academy of Engineering and Canon Medical Research Europe. This work was partially supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. We thank Nvidia for donating a Titan-X GPU. S.A. Tsaftaris acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819\8\25).

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[3] Olivier Bernard, Alain Lalande, Clement Zotti, and et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging (TMI)*, 37(11):2514–2525, 2018.

[4] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[5] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(3):740–757, 2019.

[6] Xiaobin Chang, Tao Xiang, and Timothy M Hospedales. Scalable and effective deep cca via soft decorrelation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1488–1497, 2018.

[7] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. Domain adaptation for upper body pose tracking in signed TV broadcasts. In *Proc. British Machine Vision Conference (BMVC)*, 2013.

[8] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58, 2019.

[9] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in VAEs. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 2615–2625, 2018.

[10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing

generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 2172–2180, 2016.

[11] Taco S. Cohen and Max Welling. Learning the irreducible representations of commutative lie groups. In *Proc. International Conference on Machine Learning (ICML)*, pages 1755–1763, 2014.

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[13] Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz S Kandola. On kernel-target alignment. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 367–373, 2002.

[14] Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 4414–4423, 2017.

[15] Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. In *arXiv preprint arXiv:1210.5474*, 2012.

[16] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[17] Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2020.

[18] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2018.

[19] Patrick Esser, Johannes Haux, and Bjorn Ommer. Unsupervised robust disentangling of latent characteristics for image synthesis. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2699–2709, 2019.

[20] Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *International Conference on Learning Representations (ICLR)*, 2020.

[21] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proc. International conference on algorithmic learning theory (ALT)*, pages 63–77. Springer, 2005.

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017.

[23] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

[24] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[25] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 517–526, 2018.

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017.

[27] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 179–196, 2018.

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[29] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proc. International Conference on Machine Learning (ICML)*, pages 2649–2658, 2018.

[30] Tejas D. Kulkarni, William F. Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, page 2539–2547, 2015.

[31] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations (ICLR)*, 2018.

[32] Gihyun Kwon and Jong Chul Ye. Diagonal attention and style-based GAN for content-style disentanglement in image generation and translation. *arXiv preprint arXiv:2103.16146*, 2021.

[33] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proc. European Conference on Computer Vision (ECCV)*, pages 36–52, 2018.

[34] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 700–708, 2017.

[35] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016.

[36] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Learning Representations Workshops (ICLRW)*, 2019.

[37] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A commentary on the unsupervised learning of disentangled representations. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pages 13681–13684, 2020.

[38] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019.

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019.

[40] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 7198–7211, 2020.

[41] Esser Patrick, Ekaterina Sutter, and Björn Ommer. A variational U-Net for conditional appearance and shape generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8857–8866, 2018.

[42] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[43] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *Proc. International Conference on Machine Learning (ICML)*, pages 1431–1439, 2014.

[44] Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the F-statistic loss. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, page 185–194, 2018.

[45] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016.

[46] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. ALADIN: All layer adaptive instance normalization for fine-grained style similarity. *arXiv preprint arXiv:2103.09776*, 2021.

[47] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations (ICLR)*, 2020.

[48] N. Siddharth, B. Paige, J.-W. van de Meent, A. Desmaison, N. D. Goodman, P. Kohli, F. Wood, and P. Torr. Learning disentangled representations with semi-supervised deep generative models. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[49] Zengjie Song, Oluwasanmi Koyejo, and Jiangshe Zhang. Toward a controllable disentanglement network. *arXiv preprint arXiv:2001.08572*, 2020.

[50] Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Royal Danish Academy of Sciences and Letters*, 5(4): 1–34, 1948.

[51] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

[52] Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? on the generalization of representations learned from correlated data. *arXiv preprint arXiv:2006.07886*, 2020.

[53] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[54] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *International Conference on Learning Representations (ICLR)*, 2017.

[55] Yijun Xiao and William Yang Wang. Disentangled representation learning with Wasserstein total correlation. *arXiv preprint arXiv:1912.12818*, 2019.

[56] Xianglei Xing, Tian Han, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Unsupervised disentangling of appearance and geometry by deformable generator network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10354–10363, 2019.

[57] Jimei Yang, Scott E. Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 1099–1107, 2015.

[58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[59] Sharon Zhou, Eric Zelikman, Fred Lu, Andrew Y. Ng, Gunnar Carlsson, and Stefano Ermon. Evaluating the disentanglement of deep generative models through manifold topology. In *International Conference on Learning Representations (ICLR)*, 2021.

[60] Xinqi Zhu, Chang Xu, and Dacheng Tao. Where and what? examining interpretable disentangled representations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.